搜狐携手新一代英特尔® 至强® 可扩展处理器 提升 AI 推理性能 加速负载均衡加解密处理

“第三代英特尔® 至强® 可扩展处理器带来了可观的性能提升，以及领先的技术特性，这支持我们在不大幅改变现有基础设施架构的前提下，显著提升应用系统的性能，同时提高投资回报。我们将与英特尔进一步探索在更多领域的合作，以融合双方的创新能力，构建下一代的数据中心基础设施，在性能、敏捷性、扩展性等方面实现更大的优化。”

王帅

搜狐大数据中心总经理

概述

作为一个由数字技术创新驱动、竞争激烈的行业，互联网借助视频、资讯、音乐、游戏、搜索等丰富多彩的服务，促进了物理世界与虚拟世界的交融。为了给用户带来更加丰富、个性化、响应及时的互联网服务，互联网厂商借助人工智能（AI）、大数据等技术，开展了新一轮数字化创新之旅。对于后端的数据中心基础设施来说，这些技术创新带来了对于算力、内存与存储、网络等关键资源的强劲需求，也带来了较高的扩展性、经济性等方面的压力。互联网用户普遍强化了新一代硬件的部署与软件优化，以加速构建面向未来的数据中心基础设施。

中国领先的互联网媒体、娱乐、在线游戏集团搜狐已经成为拥有诸多知名产品的超级互联网平台，秉承“让网络成为中国人民生活中不可缺少的一部分”的理想，搜狐在技术创新上进行广泛投入，推动数据中心基础设施的持续进化。在此过程中，搜狐与英特尔实现了紧密的合作，在基于 AI 的搜狐推荐业务系统中采用了第三代英特尔® 至强® 可扩展处理器与英特尔® OpenVINO™ 工具套件，显著提升了 AI 推理性能。此外，搜狐还在 Nginx 服务器中使用了英特尔® 密码操作硬件加速指令（英特尔® Crypto-NI）加速加解密负载，有效化解了 Web 业务的算力瓶颈。

背景：技术创新凸显数据中心基础设施的算力瓶颈

受益于数字化技术创新，用户能够在互联网上体验到更加丰富多彩的服务。在本轮数字化创新浪潮中，值得关注的重要趋势包括：AI 更加广泛地变革了互联网数据的处理模式以及互联网业务的交付模式，智能化成为互联网企业竞相发力的竞争点，智能搜索、AI 推荐等应用让用户能够得到更加个性化的应用体验；网络与数据安全的重要性不断凸显，保护用户数据安全不仅是合规性强制要求，同时也在很大程度上影响了业务的连续性与企业商誉；业务创新的频率加快，并更加强调依托开源生态，进行快速开发、快速测试、快速部署与持续迭代。

随着这些技术创新的不断落地，以及互联网业务数据的爆炸式增长，数据中心基础设施面临着算力、扩展性、敏捷性、经济性等多方面的挑战。其中，尤为值得关注的是算力的巨大缺口。据 IDC 统计，近 10 年来全球算力增长明显滞后于数据增长。全球算力的需求每 3.5 个月就会翻一倍，远远超过了当前算力的增长速度。对于互联网企业来说，要弥补算力缺口，同时提升数据中心基础设施建设的投资回报，可以着重从以下几个角度来思考：

- **如何利用新一代硬件来提升数据中心的资源密度：**遵循摩尔定律，CPU 等硬件性能实现了持续的代际提升，在性能/成本上有着更大的优势，而且有利于提供更高的资源密度，这将有助于企业降低在数据中心建设、运维、能耗等方面的成本，同时为 AI 等新型负载提供更佳的算力支撑。
- **如何通过软件优化以挖掘基础设施潜能：**对于 AI、加解密等负载来说，软件优化对于性能提升有着重要的意义。以 AI 为例，通过选择经过预训练、优化的高效 AI 模型，将模型量化至低精度，生成最小化运行时软件包等方式，能够有效提升 AI 模型的推理速度，降低在专用硬件方面的需求。
- **如何利用异构化硬件以支持多样化负载：**互联网企业的数据中心的负载正在日趋多样化，用户将需要通过 CPU、XPU 等芯片构建创新的基础设施，软件将作为微服务进行部署，以适应计算、存储和内存不断分离的新特征。
- **如何通过全面的数据中心组合来实现性能与负载优化：**现代化的互联网数据中心正在变得日趋复杂化，通过涵盖 CPU、GPU、持久内存和存储技术、以太网适配器、FPGA、优化的软件解决方案、AI 加速器、硅光子学，以及增强的内置安全功能在内的全面数据中心组合，可在所有平台上实现性能和工作负载优化。

聚焦新技术、新商业下的新趋势，搜狐显著强化 AI、5G 等技术创新，搜狐媒体平台和视频平台丰富优质的内容，与智能技术结合，正在实现从内容到品牌、从 IP 质感到产品质感的无缝传导，并提供人性化的智能营销体系，通过可预测、动态化、感知情绪和感知场景的科技，帮助品牌理解用户的真实意图和情绪状态，让广告与用户的需求和场景呼应。同时，搜狐与英特尔等伙伴合作，在云数据中心方面进行了广泛创新，为前端业务提供了澎拜的基础资源。

解决方案：搜狐与英特尔协同进行 AI 推荐与 HTTPs 性能优化

在数字经济时代，搜狐与英特尔构建了密切的合作关系，并围绕新一代硬件部署、软件优化等方面进行了长期的协同创新，获得了大量的成果。本白皮书主要展示了双方在 AI 推荐与 HTTPs 性能优化方面的合作实例：

AI 推荐性能优化

为了整合全站的视频、资讯等资源，将个性化、新鲜的互联网资源快速分发给不同应用场景的用户，提升用户体验，搜狐推出了基于 AI 技术的推荐系统。该系统包括知识库、主题模型、用户/视频画像、实时反馈/统计、独立后台、推荐引擎、视频处理引擎等基础组件，能够对于亿级的海量数据进行分析，并通过实时特征工程、在线学习、多模型融合等技术进行智能排序。在该 AI 推荐引擎中，搜狐主要采用基于 gRPC 通信方式的各种网络结构的深度模型。

为了充分利用现有的基础设施，控制 AI 推荐系统的总体拥有成本（TCO），搜狐采用了基于 CPU 的 AI 推荐方案。在该方案中，搜狐采用了基于第三代英特尔® 至强® 可扩展处理器的推荐服务器，以尽可能提升推荐服务器的性能表现。

图 1. 搜狐已成为拥有诸多知名产品的超级互联网平台

和上一代产品相比，第三代英特尔® 至强® 可扩展处理器在性能和支持的内容容量方面均有显著提高，并且具备一系列特性以支持各种复杂的工作负载，有助于推动经济高效、灵活且可扩展的数据中心计算架构，为 AI、数据分析等关键任务提供增强的每节点性能。第三代英特尔® 至强® 可扩展处理器内置了英特尔® 深度学习加速技术，该技术在指令集中新增了英特尔® AVX-512 VNNI（矢量神经网络指令集），后者是对标准英特尔® AVX-512 指令集的扩展。英特尔® AVX-512 VNNI 将三条指令合并成一条指令执行，可更充分地发挥新一代英特尔® 至强® 可扩展处理器的计算潜能，提升 INT8 模型的推理性能。

在工作中，未使用 VNNI 的平台需要 vpmaddubsw、vpmaddwd 和 vpaddd 指令才能完成 INT8 卷积运算中的乘累加：

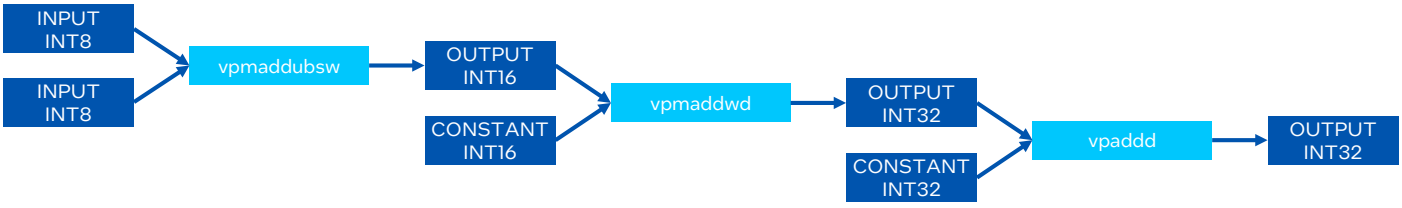

图 2-1. 未使用 VNNI 的 INT8 卷积运算流程

而拥有 VNNI 的平台只需使用一条指令 vpdpbusd 即可完成 INT8 卷积操作：

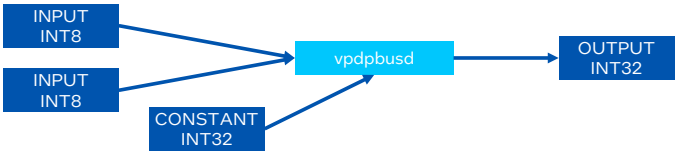

图 2-2. 使用 VNNI 的 INT8 卷积运算流程

为了进一步加速 AI 推荐系统的性能表现，搜狐还采用了 OpenVINO™ 工具套件对其进行优化。OpenVINO™ 工具套件是用于快速开发应用程序和解决方案，以解决各种任务（包括人类视觉模拟、自动语音识别、自然语言处理和推荐系统等）的综合工具套件。该工具套件基于新一代的人工神经网络，包括卷积神经网络（CNN）、递归网络和基于注意力的网络，可跨英特尔® 硬件扩展计算机视觉和非视觉工作负载，从而大幅提高性能。

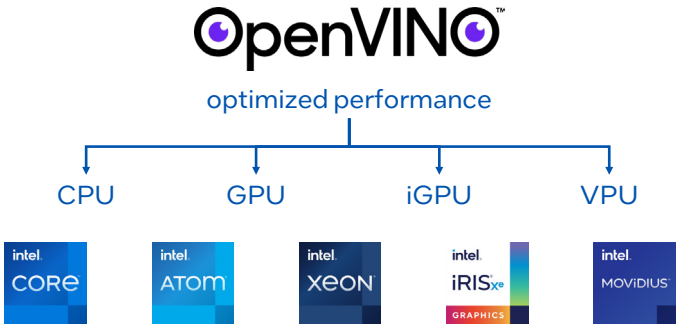

图 3. OpenVINO™ 工具套件支持的硬件组合

通过在基于第三代英特尔® 至强® 可扩展处理器的推荐服务器上使用 OpenVINO™ 工具套件进行加速，搜狐显著提升了 DeepFM 模型的推理性能。测试数据显示，在 Tensorflow 下，运行于英特尔® 至强® 金牌 6330 处理器之上的 Deepfm 模型时延相较于英特尔® 至强® E5-2650v4 处理器的时延，降低接近 50%[1]。在 OpenVINO™ 工具套件下，运行于英特尔® 至强® 金牌 6330 处理器之上的 ResNet50 模型带宽相较于英特尔® 至强® E5-2650v4 有高达 3.4 倍[2] 的提升。

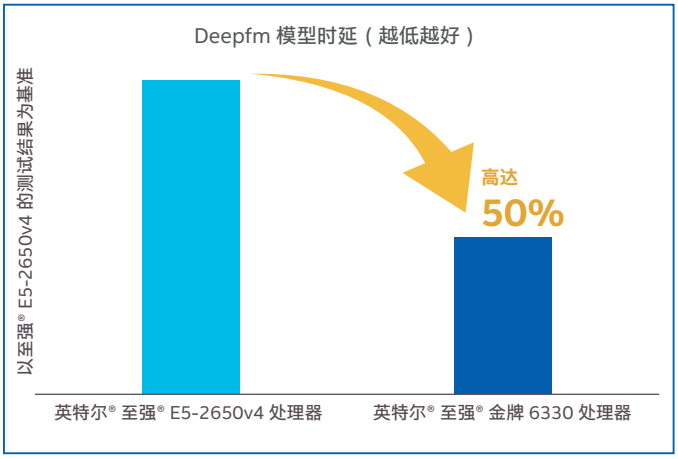

图 4. Deepfm 时延测试[3]

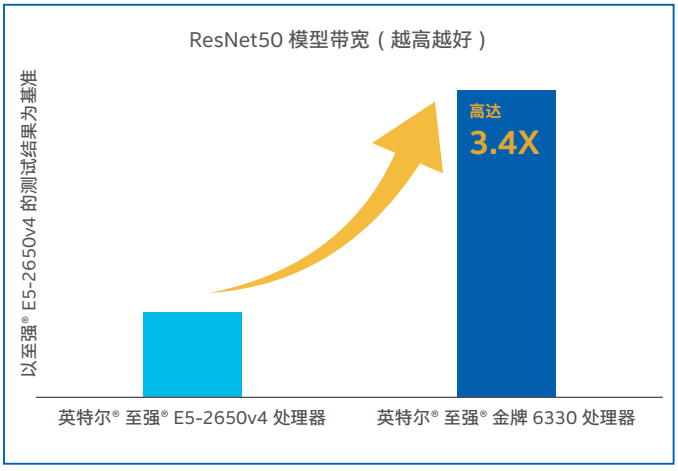

图 5. ResNet50 带宽测试[4]

在 Tensorflow下使用 gRPC 模型的测试显示，在基于英特尔® 至强® E5-2650v4 的测试平台上，经过 OpenVINO™ 工具套件优化后，模型性能有了 2 倍的提升。而在将平台的处理器升级为英特尔® 至强® 金牌 6330 处理器之后，模型性能有高达 3.4 倍[5] 的提升。

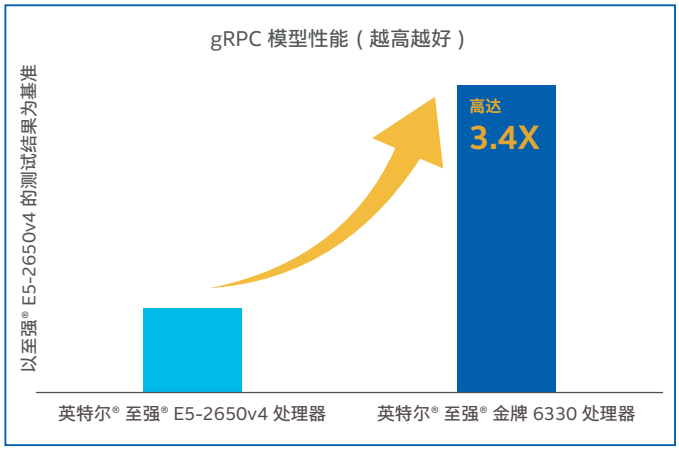

图 6. gRPC 模型性能测试[6]

HTTPs 性能优化

HTTPs（Hyper Text Transfer Protocol over Secure Socket Layer）作为 HTTP 的有效安全升级，已成为互联网安全接入的重要趋势。为了提高互联网业务与数据的安全性，搜狐将旗下的业务大规模迁移至 HTTPs 加密传输。但与此同时，HTTPs 在提升安全性的同时，也会在性能与 TCO 方面带来巨大压力。

HTTPs 的性能瓶颈在很大程度上是因为，HTTPs 的安全基础是安全传输层（Transport Layer Security，TLS）/ 安全套接字协议（Secure Sockets Layer，SSL）。而 TLS/SSL 协议提供的连接安全性具有私有、可靠性两大特点，其利用非对称加密算法实现身份认证和密钥协商，对称加密算法采用协商的密钥对数据加密，基于散列函数验证信息的完整性。HTTPs 带来的数据加解密负载将会消耗大量的 CPU 资源，在同等硬件条件下，未经优化的 HTTPs 业务可能会比 HTTP 高出数百毫秒，影响用户访问体验。

在 HTTPs 业务中，搜狐使用了 Nginx 服务器。Nginx 是一个高性能的 HTTP 和反向代理 Web 服务器，同时也提供了 IMAP/POP3/SMTP 服务，通过启用异步模式，Nginx 能够通过并行处理减少等待，消耗相对很少的系统资源就能达到所需的性能，缩短应用响应时间。

要解决 Nginx 服务中的加解密性能问题，主要有两大方式。一种是将加解密负载卸载到英特尔® QAT 等专用的加速卡上，另一种是完全依赖 CPU 进行处理，这种方式在相当长的时间内缺乏足够的效率，但是随着搭载英特尔® Crypto-NI 的第三代英特尔® 至强® 可扩展处理器的推出，其实用性开始凸显出来。

[1,2,3,4,5,6] 数据援引自搜狐于 2022 年 6 月开展的测试。基准配置：单节点，2x 英特尔® 至强® E5 2650v4 处理器，256 GB 总内存，超线程：启用，睿频：启用，存储（启动）：1x 400 GB DC3700，存储（应用）：2 * 4TB DC P4500 PCIe NVME，网络设备：2 x 82599ES 双端口 10GbE，网络速度：10GbE，操作系统：Red Hat Enterprise Linux* 7.4，Kernel: 3.10.0-693.11.6.el7.x86_64 x86_64。新配置：单节点，2x 英特尔® 至强® 金牌 6330 处理器，256 GB 总内存，超线程：启用，睿频：启用，存储（启动）：1x 400 GB DC3700，存储（应用）：2 * 4TB DC P4500 PCIe NVME，网络设备：2 x 82599ES 双端口 10GbE，网络速度：10GbE，操作系统：Red Hat Enterprise Linux* 7.4，Kernel: 3.10.0-693.11.6.el7.x86_64 x86_64。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

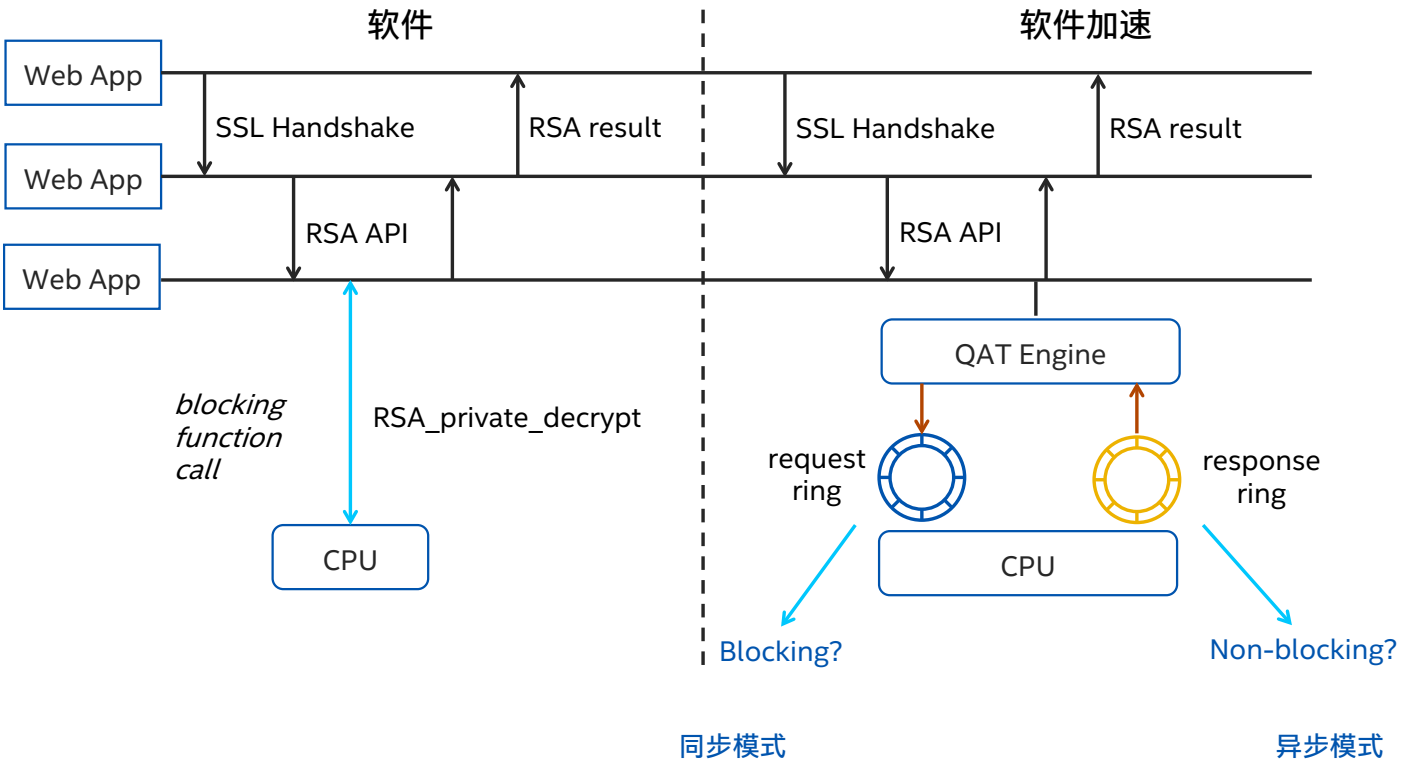

图 7. 英特尔® Crypto-NI 加速示意

英特尔® Crypto-NI 加入了一系列基于英特尔® AVX-512 的全新 SIMD（Single Instruction Multiple Data，单指令多数据流）ISA（指令集架构），如 VPMADD52 等，从硬件层面对 HTTPs 等涉及的加解密计算负载提供加速，且新处理器平台也利用英特尔 AVX-512，提供全新的英特尔® Crypto-多缓冲区软件库。

英特尔® Crypto-多缓冲区软件库融合了基于英特尔® AVX-512 指令集的 IFMA（整数融合乘法加法）指令，能针对 RSA 等算法提供深层次的软件优化。当 HTTPs 等进程使用这一技术时，可通过批处理方式对队列中维护的多个请求执行操作，并使用 OpenSSL 异步基础架构将最多 8 个批处理请求提交给多缓冲区 API，利用英特尔® AVX-512 指令集对其进行并行处理。这一多缓冲区优化的方式在具有许多并行连接的异步操作场景中使用时，可获得巨大的性能提升。

测试结果显示，相较于 Crypto NI 开启前，Crypto NI 开启后 Encryption Sign 的性能可提升高达 5 倍，Nginx QPS 可提升高达 2.6 倍[7]。

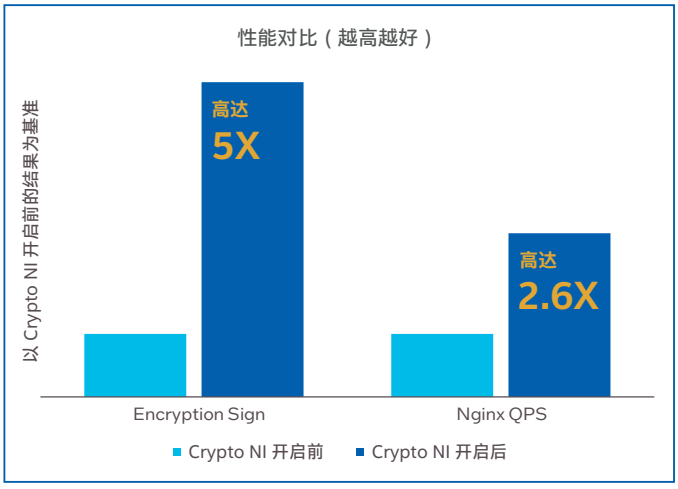

图 8. Crypto NI 开启前后的性能对比[8]

[7,8] 数据援引自搜狐于 2022 年 6 月开展的测试。测试配置：单节点，2x 英特尔® 至强® 金牌 6330 处理器，256 GB 总内存，超线程：启用，睿频：启用，存储（启动）：1x 400 GB DC3700，存储（应用）：2 * 4TB DC P4500 PCIe NVME，网络设备：2 x 82599ES 双端口 10GbE，网络速度：10GbE，操作系统：Red Hat Enterprise Linux* 7.4，Kernel: 3.10.0-693.11.6.el7.x86_64 x86_64。英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

收益：数据中心基础设施优化提升业务价值

搜狐与英特尔在 AI 推荐性能优化与 HTTPs 性能优化方面的合作证实，通过采用融合了众多创新特性的新一代的英特尔® 至强® 可扩展处理器，并结合 OpenVINO™ 工具套件等软件工具，能够显著提升面向 AI、加解密等负载的应用系统性能，并保证足够的敏捷性与扩展性。

具体来说，搜狐与英特尔在上述两个解决方案上的合作带来了如下收益：

高性能

基于第三代英特尔® 至强® 可扩展处理器以及 OpenVINO™ 工具套件的优化，支持搜狐在现有基础设施上对于性能进行充分优化，从多个层面大幅提升了相关计算性能，满足了上层应用对于性能的强劲需求。

高经济性

该方案在硬件方面主要依赖于 CPU，无需在 GPU、加速卡等硬件上进行专门的投资。同时，服务器的能耗、空间占用等成本也得到降低，实现了更高的经济性，有助于提升投资回报率。

高稳定性

得益于架构以及硬件的简化，搜狐能够有效降低应用系统的故障点，降低运维压力，这有助于提升系统的稳定性与可用性。

展望：软硬件全栈创新释放数据潜能

不断变化的业务需求大规模加速了云技术和基础设施的采纳和部署，AI、大数据、万物互联等应用带来了指数级增长的算力需求。在此背景下，企业的云与数据中心战略面临着业务弹性、成本优化、工作负载策略、云迁移、安全等方面带来的综合挑战。英特尔推出了面向未来数据中心的异构参考架构，包含用于分布式智能的跨产品组合硬件和软件架构，可以全面提升未来数据中心的规模和效率，实现从云端到边缘端、跨越整个数据管道的解决方案。

近年来，英特尔与搜狐构建了紧密的合作，在软件定义存储性能优化、语音识别应用、推荐系统的加速与优化、DPDK 网络负载均衡加速等领域进行携手，利用新一代软硬件技术突破性能瓶颈、释放业务潜能，以全栈技术驱动搜狐的业务创新。未来，双方还计划进一步强化面向未来数据中心的异构参考架构设计与优化，进一步加速技术创新与业务价值挖掘。

关于搜狐

搜狐是中国领先的互联网媒体、娱乐、在线游戏集团，凭借强大的竞争实力，搜狐已经发展成为拥有诸多知名产品的超级互联网平台。目前，搜狐已经初步实现了从创立伊始确立的“让网络成为中国人民生活中不可缺少的一部分”的理想。在中国网民呈现爆发式增长的过程中，搜狐也始终在为大多数中国网民提供优质服务。

关于英特尔

英特尔（NASDAQ: INTC）作为行业引领者，创造改变世界的技术，推动全球进步并让生活丰富多彩。在摩尔定律的启迪下，我们不断致力于推进半导体设计与制造，帮助我们的客户应对最重大的挑战。通过将智能融入云、网络、边缘和各种计算设备，我们释放数据潜能，助力商业和社会变得更美好。如需了解英特尔创新的更多信息，请访问英特尔中国新闻中心 newsroom.intel.cn 以及官方网站 intel.cn。